Plant science data management and integration: the heterogeneity and dispersion challenge

Solutions from ELIXIR and EMPHASIS European Infrastructure and beyond

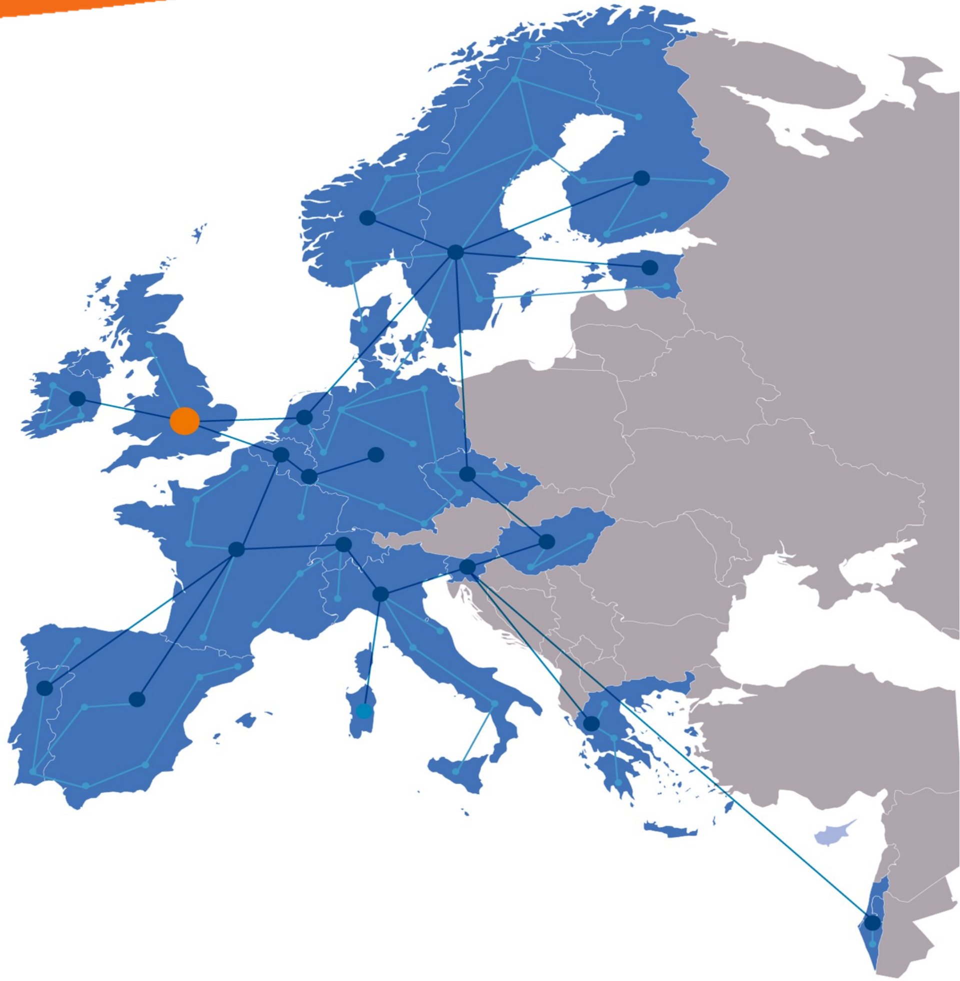
www.elixir-europe.org

ELIXIR

- intergovernmental organisation
- life science resources
  - databases, software tools, training materials, standards, compute resources
  - across Europe.

coordinate life science resources ➔ single infrastructure:

- Find and share data
- Exchange expertise
- Agree on best practises in scientific research



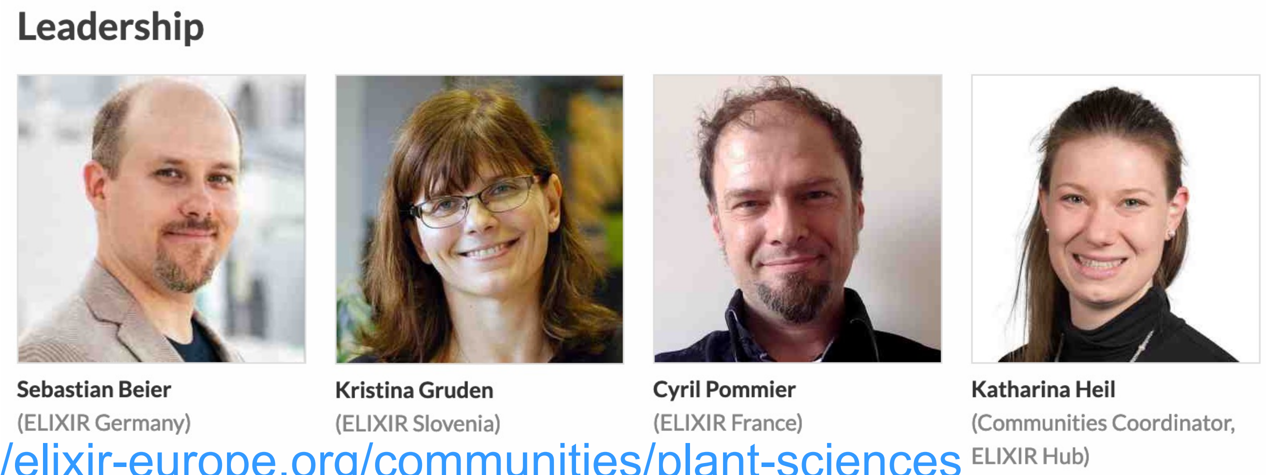
ELIXIR's work

ELIXIR coordinates activities through at least one of five areas of activities called Platforms:

- Compute
- Data
- Interoperability
- Tools
- Training

Driven by eleven **ELIXIR** Communities

Plant Sciences Community



Leadership

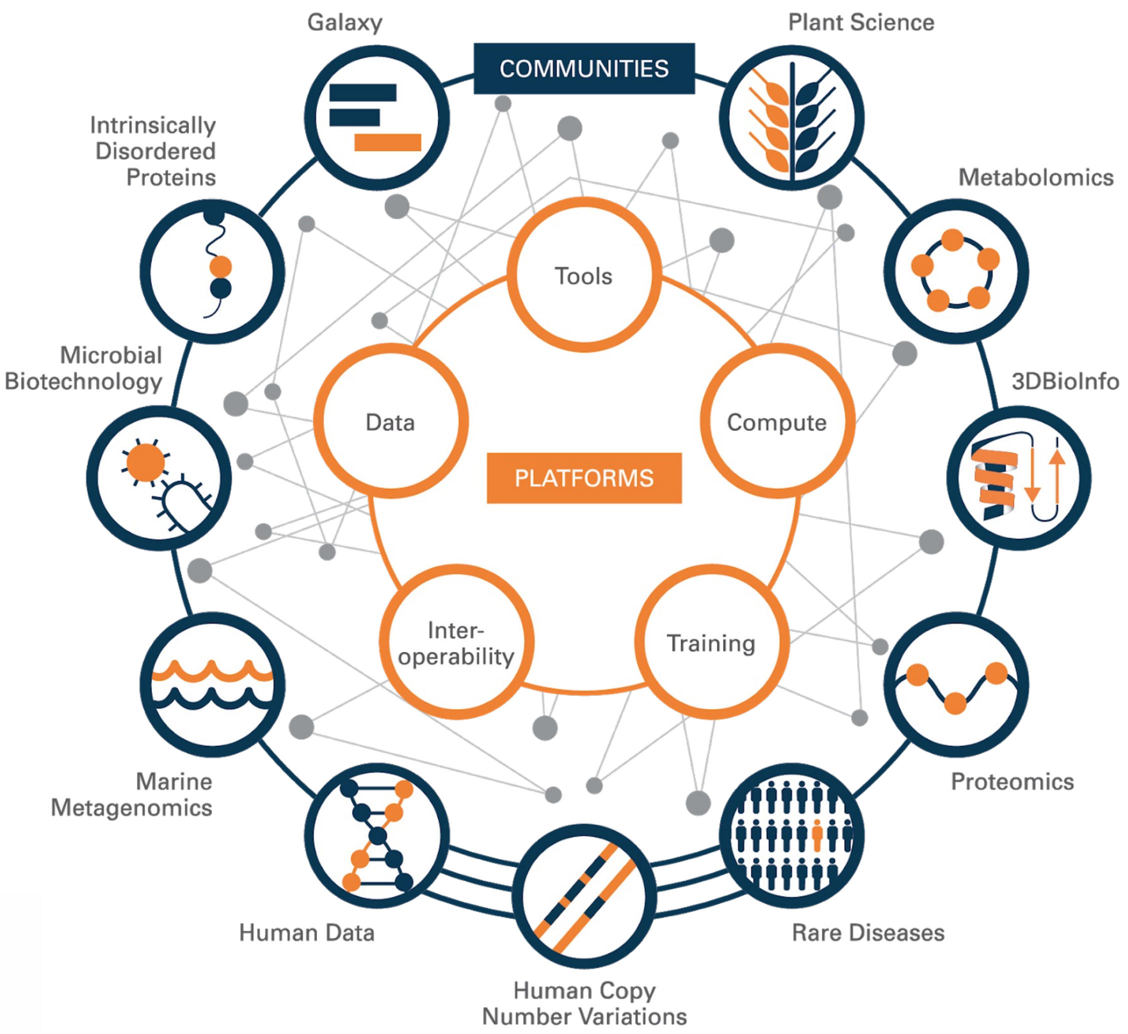
Sebastian Beier
(ELIXIR Germany)

Kristina Gruden
(ELIXIR Slovenia)

Cyril Pommier
(ELIXIR France)

Katharina Heil
(Communities Coordinator, ELIXIR Hub)

https://elixir-europe.org/communities/plant-sciences

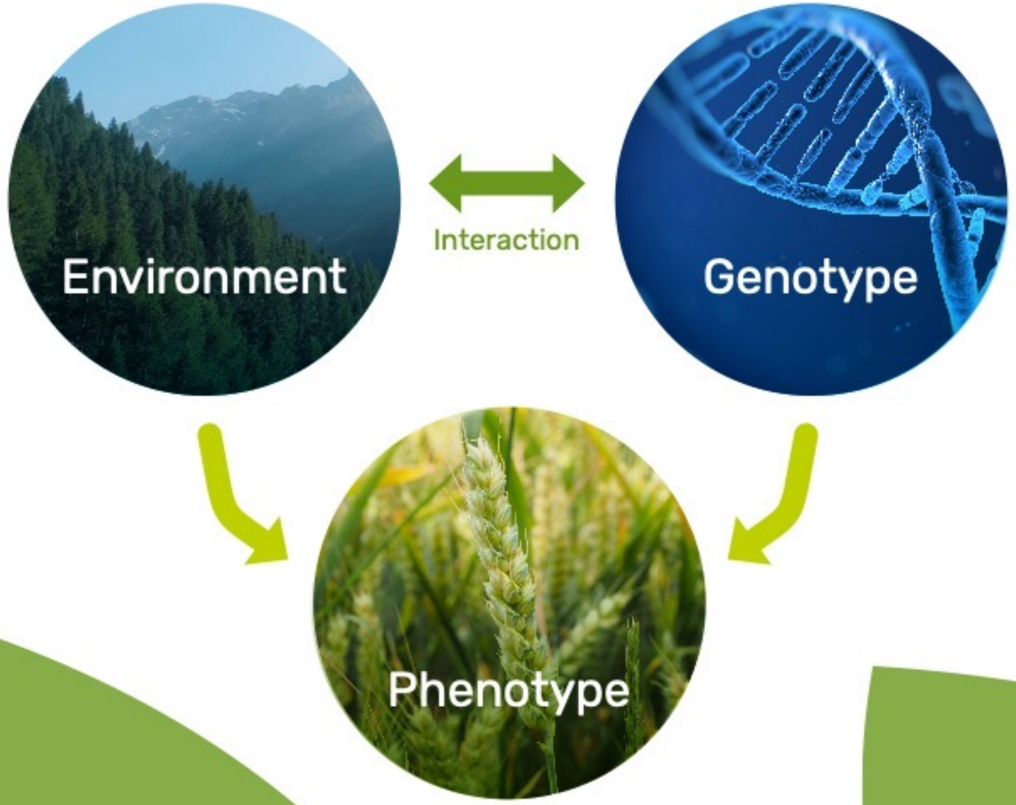


EMPHASIS

EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING

SCIENTIFIC TOOL TO STUDY PLANT–ENVIRONMENT INTERACTION

- Study of plant structure and function
- Using non-invasive technology
- Understanding how plant structure and function depend on genetics and the environment

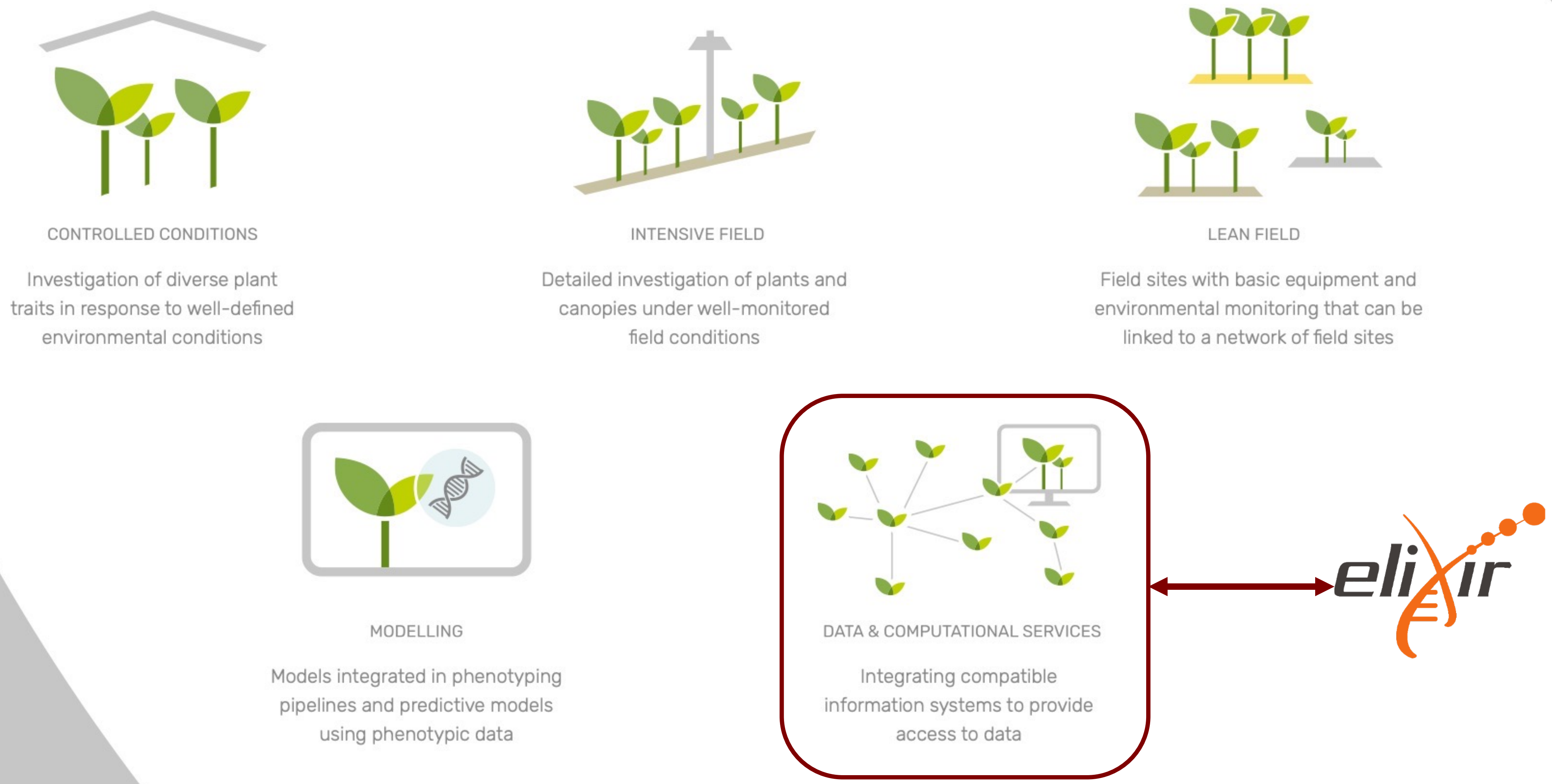


How does a plant cope with its environment?

Phenotyping is used to understand how plants can cope with reduced resources, pathogens and climate change.

Infrastructure Categories

PLANT PHENOTYPING REQUIRES INTEGRATION OF BOTH FACILITIES AND ACTIVITIES



CONTROLLED CONDITIONS

Investigation of diverse plant traits in response to well-defined environmental conditions

INTENSIVE FIELD

Detailed investigation of plants and canopies under well-monitored field conditions

LEAN FIELD

Field sites with basic equipment and environmental monitoring that can be linked to a network of field sites

MODELLING

Models integrated in phenotyping pipelines and predictive models using phenotypic data

DATA & COMPUTATIONAL SERVICES

Integrating compatible information systems to provide access to data

Open science through FAIR data principles

Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (2016)*

Findable

Ids

Index

Metadata Description

Accessible

Open Protocole

Perenial Metadata

Interoperable

Semantics

Linked Data

Vocabularies

Reusable

License

Well described

Provenance (origin, process, methodology)

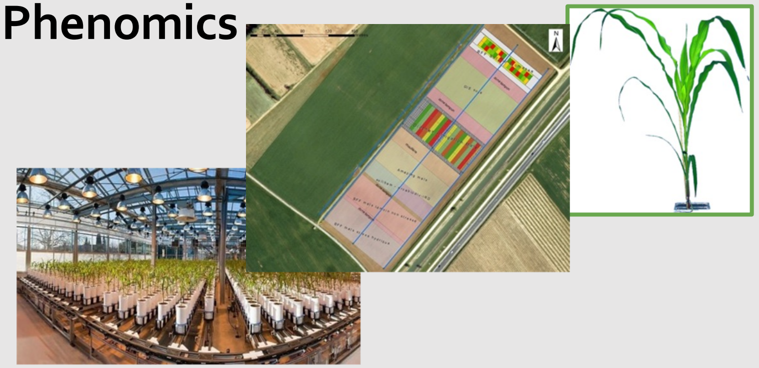
Standards

Sustainable data access over decades

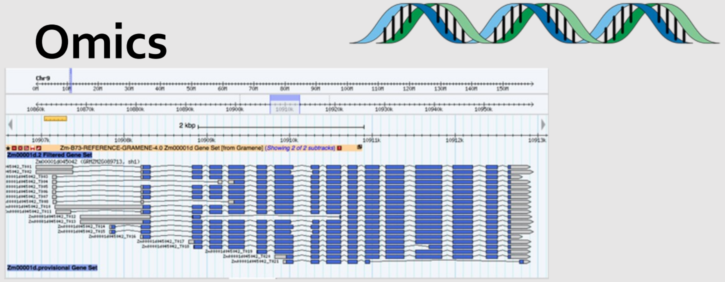
›PLANT use case

- Environment / Phenome / Genomic / *omic / Genetic

Phenomics



Genetics
Genomics
Omics



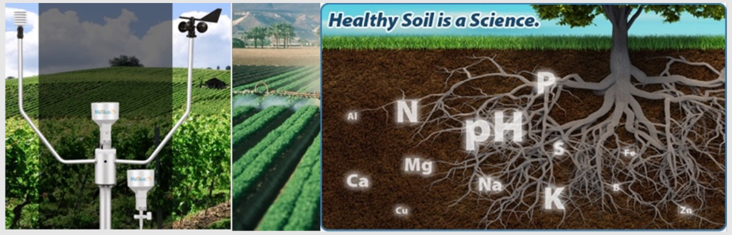
Findable Accessible Interoperable Reusable

Plant Breeding
Genetic variations by Traits

Climate Change Studie
Genotype by Environment

Findable Accessible Interoperable Reusable

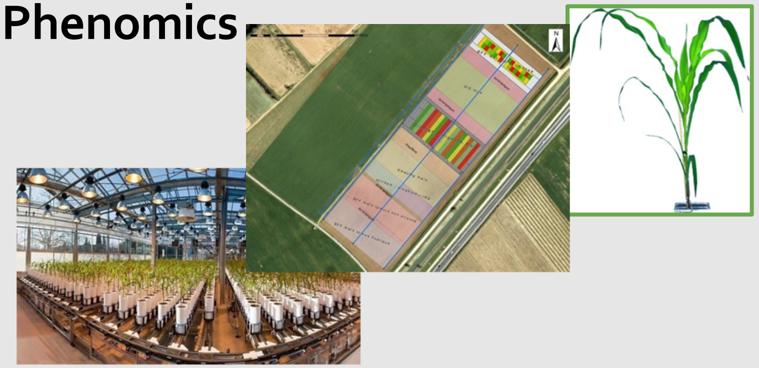
Environment



PLANT use case

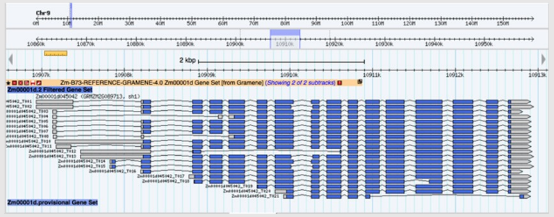
- Environment / Phenome / Genomic / *omic / Genetic

Phenomics



Dispersed
Heterogenous
Getting Standardized

Genetics
Genomics
Omics



Mostly centralized
Homogenous data
Heterogenous metadata

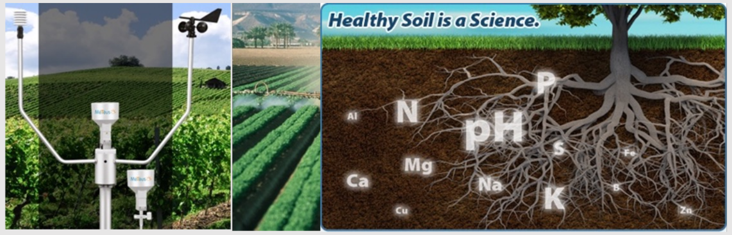
Findable Accessible Interoperable Reusable

Plant Breeding
Genetic variations by Traits

Climate Change Studie
Genotype by Environment

Findable Accessible Interoperable Reusable

Environment



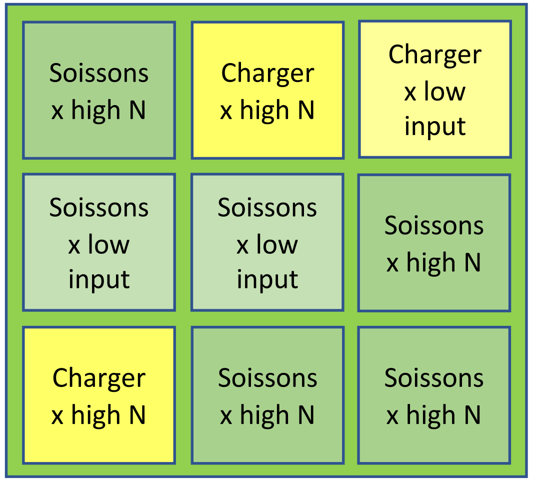
Dispersed
Heterogenous

What is FAIR for plant data ?

- Phenotyping
  - Raw data
    - Images
    - NIRS
    - Individual plant time series
    - Expensive to generate
    - Not reproducible
  - Computed / derived data
    - Data matrices (XLSX)
- Genetic variation
  - Raw data
    - Sequence files
    - “cheap” to generate
    - Big Data
  - Derived
    - VCF
    - Aligned to a given reference genome

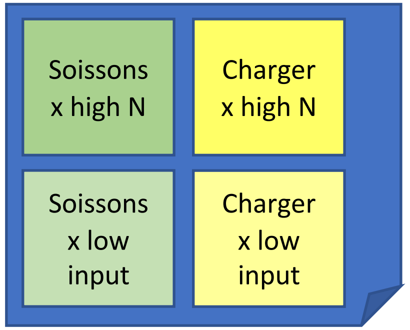
Phenotyping data life cycle

« Raw » data, pheno/env measures, variables


Soissons x high N
Charger x high N
Charger x low input
Soissons x low input
Soissons x low input
Soissons x high N
Charger x high N
Soissons x high N
Soissons x high N

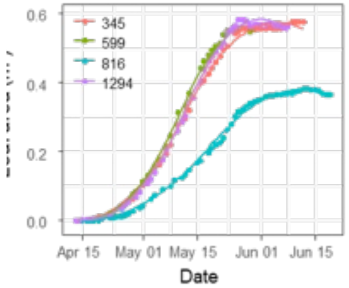
Derivation, Reduction

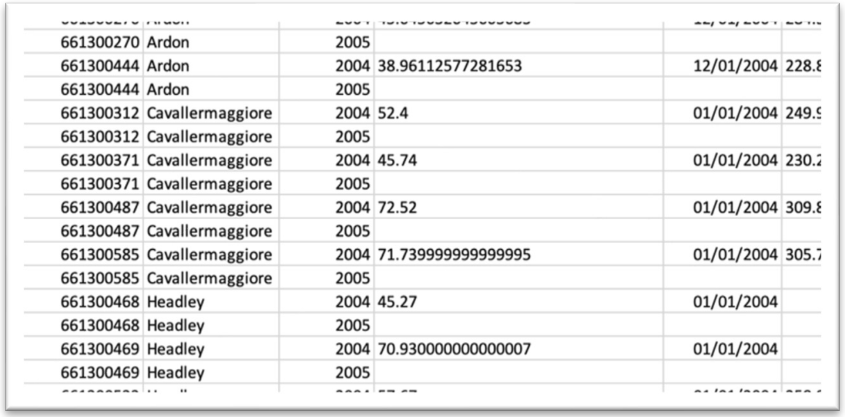
« computed » data, reduced, indicators


Soissons x high N
Charger x high N
Soissons x low input
Charger x low input

Genotype	Treatment	N input	Date	Rep	Fusariose
Soissons	low input	15,32253129	15/11/2011	1	5
Soissons	low input	15,31430556	16/11/2011	2	7

Genotype	Treatment	Fusariose
Soissons	low input	6





661300270	Ardon	2005			
661300444	Ardon	2004	38.96112577281653	12/01/2004	228.8
661300444	Ardon	2005			
661300312	Cavallermaggiore	2004	52.4	01/01/2004	249.9
661300312	Cavallermaggiore	2005			
661300371	Cavallermaggiore	2004	45.74	01/01/2004	230.2
661300371	Cavallermaggiore	2005			
661300487	Cavallermaggiore	2004	72.52	01/01/2004	309.8
661300487	Cavallermaggiore	2005			
661300585	Cavallermaggiore	2004	71.739999999999995	01/01/2004	305.7
661300585	Cavallermaggiore	2005			
661300468	Headley	2004	45.27	01/01/2004	
661300468	Headley	2005			
661300469	Headley	2004	70.930000000000007	01/01/2004	
661300469	Headley	2005			

Plant Phenotyping Life cycle

Raw data long term conservation

Data acquisition

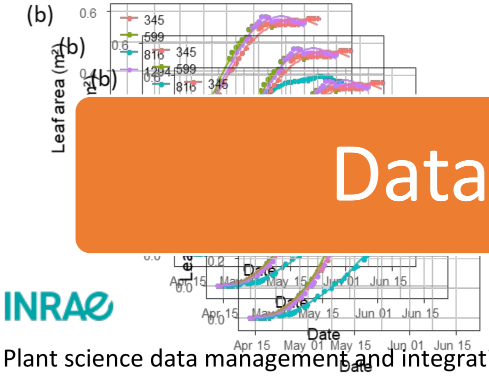
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Analysis Reproducibility
- Provenance

Data computation

- **INDICATORS**
- Statistical integration
- Genotype level (mostly)
- New computation for each scientific question
- One raw dataset → many computed datasets

Data publication

- One Data Publication by datasets.
- **Platform IS**
  - Phenomic, plant level
- **FAIR Data Repositories**
  - Reduced



Data

Genotype	traitement	Fusariose
oisson	low input	5
oisson		
Charger	low input	1
Charger	high N	2

Knowledge

Variety charger
intensiv cultural
practice

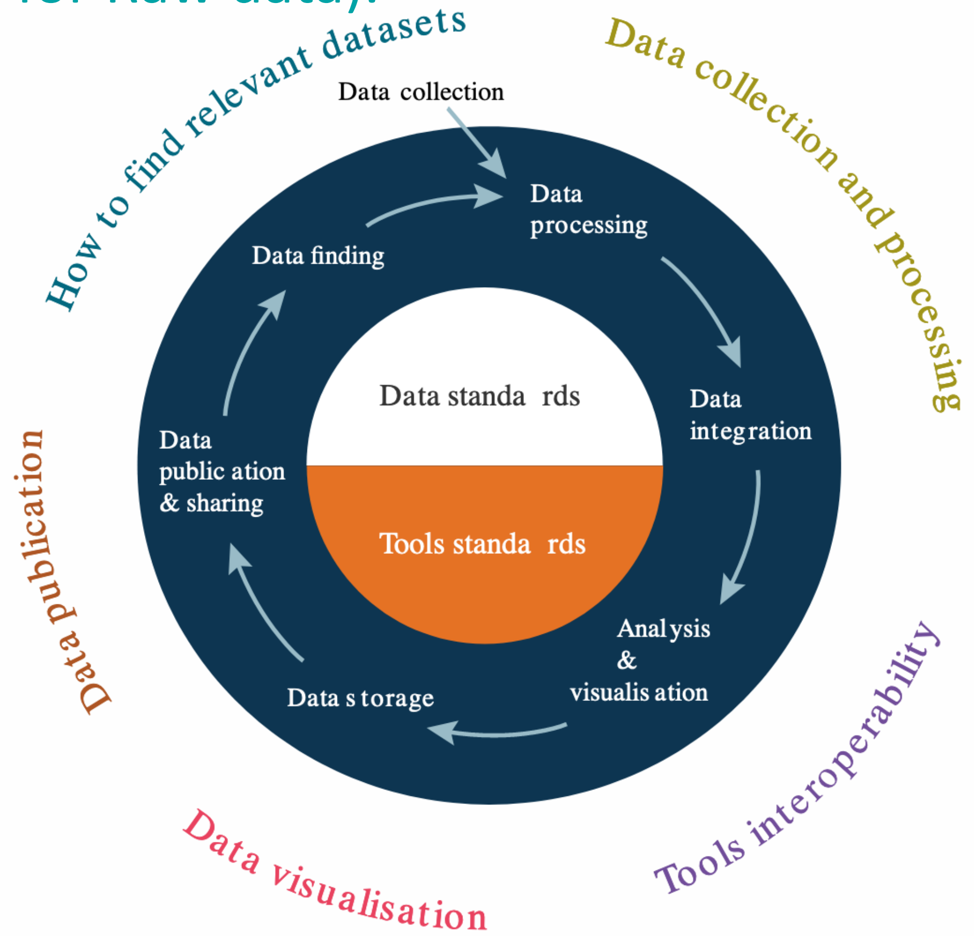
Plant Genetic variation

- Variability of the genotypes (AKA varieties, accessions, germplasm)
- Sequencing (GBS), Chips, …
- Raw data : reads
- Aligned data : VCF
- Paradigm: Raw data is too big, easy to generate ➔ keep only Variation
- But: realign to a new genome version, or to another reference variety ?
- ➔ Raw data can be interesting to keep too

FAIR For plant science

Findable Accessible Interoperable Reusable

- Phenotyping: Raw and derived data
- Genotyping: Computed data (plus option for Raw data).
  - Applies to other OMICS
- Solutions on the data lifecycle
  - Data standardisation
  - Data repositories for publication
  - Data findability / discovery

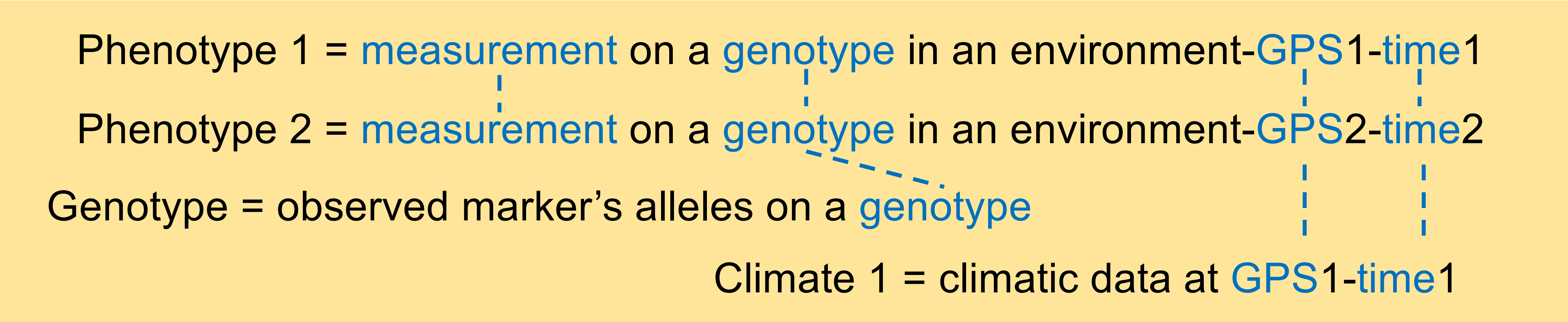


PLANT DATA STANDARDS : WHY

Why should we standardize data?

- Allow anyone (including yourself) to reuse it: metadata about the experiment (who did it, for what purpose, where and how)

- Enable data integration with other types of data: Linked data between datasets using identification of pivot objects



Phenotype 1 = measurement on a genotype in an environment-GPS1-time1

Phenotype 2 = measurement on a genotype in an environment-GPS2-time2

Genotype = observed marker’s alleles on a genotype

Climate 1 = climatic data at GPS1-time1

- To enable knowledge discovery: metadata about the experiment, controlled vocabularies, ontologies

PLANT DATA STANDARDS : WHO

Interoperability in International network

National Networks

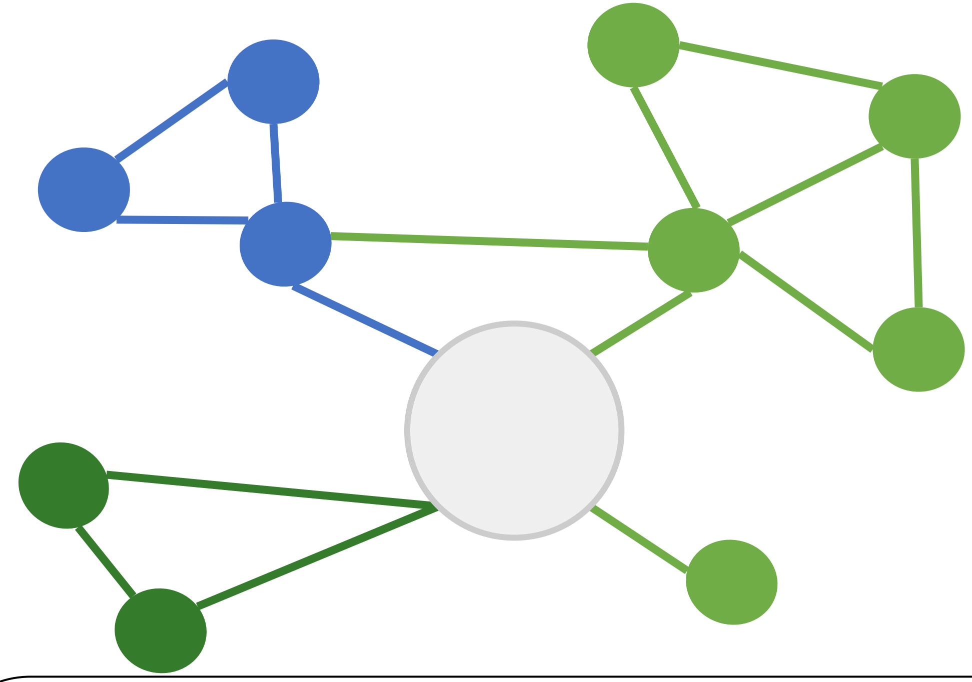
ifb de.NBI GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

PHENOME EMPHASIS French Plant Phenomic Infrastructure

Amaizing

FRANCE GRILLES

BreedWheat



European Networks

elixir

EMPHASIS

Global Networks

WheatIS

WHEAT INITIATIVE

CGIAR

International data standards

BrAPI

Web services

RDA RESEARCH DATA ALLIANCE

miappe

Minimal information

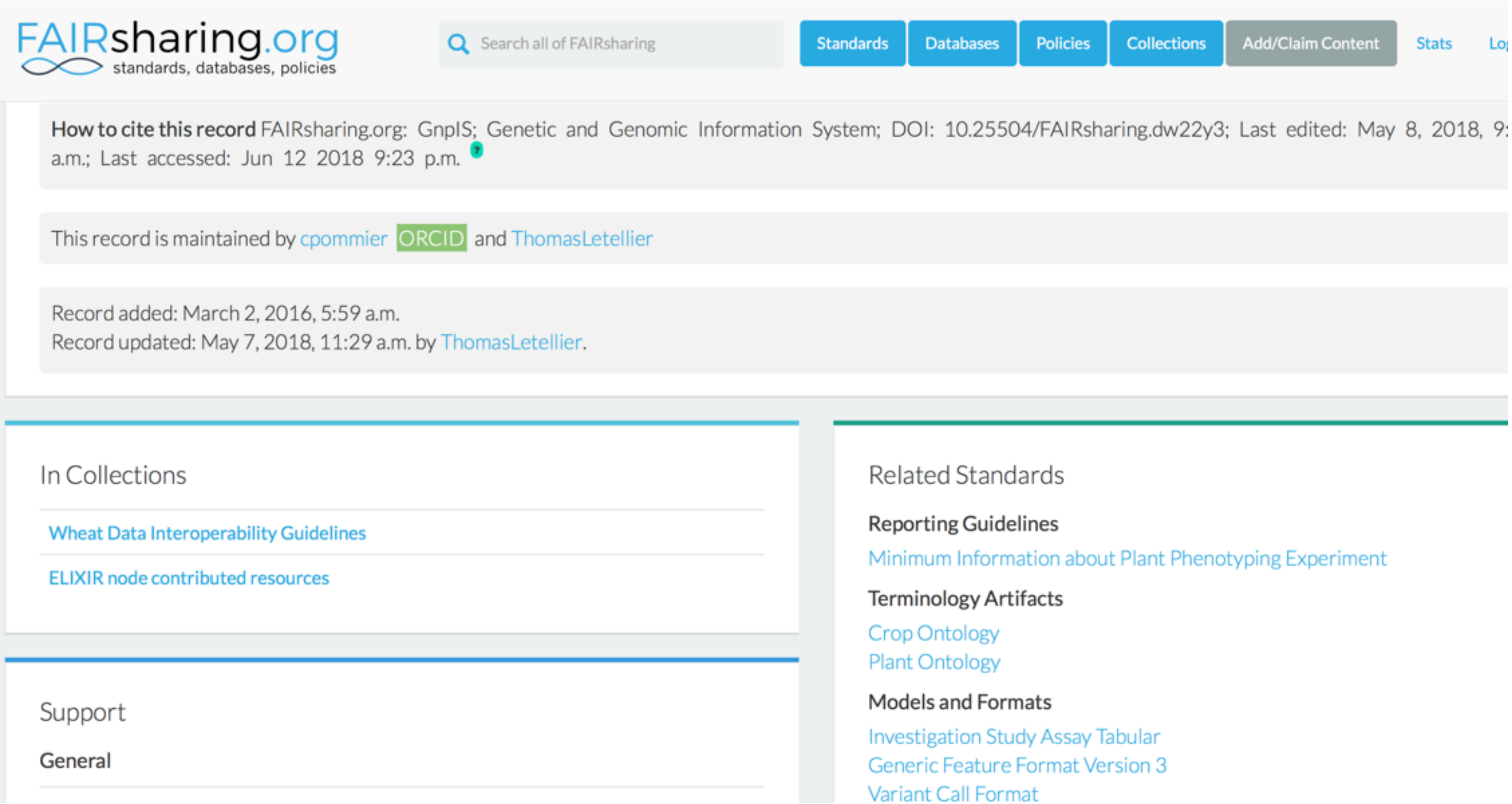
Crop Ontology for agricultural data

Controlled vocabularies
Trait dictionaries

FAO Bioversity International

MCPD

Sharing standards: standards registries



How to cite this record FAIRsharing.org: GnpIS; Genetic and Genomic Information System; DOI: 10.25504/FAIRsharing.dw22y3; Last edited: May 8, 2018, 9: a.m.; Last accessed: Jun 12 2018 9:23 p.m.

This record is maintained by cpommier ORCID and ThomasLetellier

Record added: March 2, 2016, 5:59 a.m.
Record updated: May 7, 2018, 11:29 a.m. by ThomasLetellier.

In Collections

- Wheat Data Interoperability Guidelines
- ELIXIR node contributed resources

Support

General

Related Standards

Reporting Guidelines
Minimum Information about Plant Phenotyping Experiment

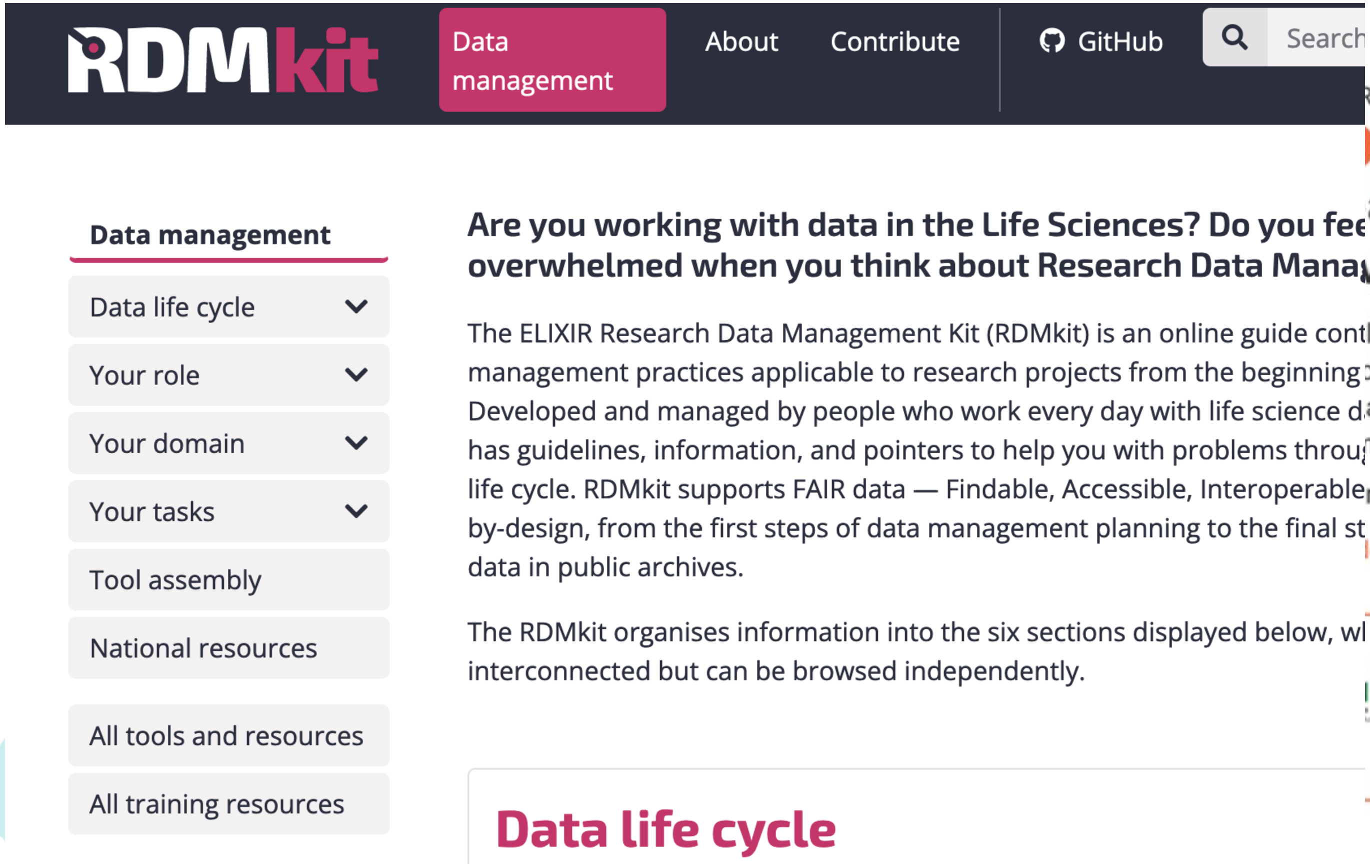
Terminology Artifacts
Crop Ontology
Plant Ontology

Models and Formats
Investigation Study Assay Tabular
Generic Feature Format Version 3
Variant Call Format

Community driven recomendations and registries

- WheatIS: http://wheatis.org/DataStandards.php
- RDM Toolkit
  - https://rdmkit.elixir-europe.org/

- Community story

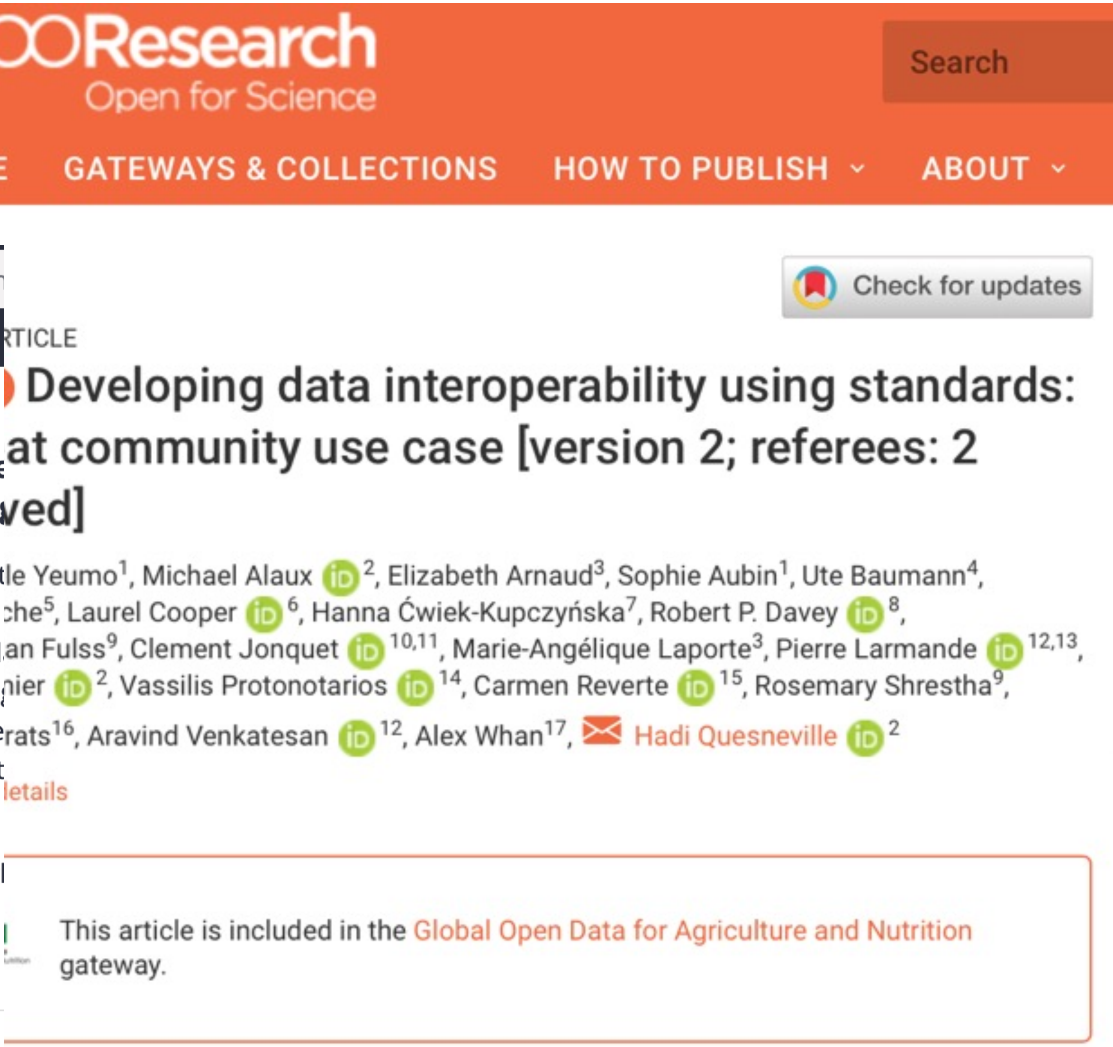


Are you working with data in the Life Sciences? Do you fee overwhelmed when you think about Research Data Mana

The ELIXIR Research Data Management Kit (RDMkit) is an online guide con management practices applicable to research projects from the beginning Developed and managed by people who work every day with life science d has guidelines, information, and pointers to help you with problems throu life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable by-design, from the first steps of data management planning to the final st data in public archives.

The RDMkit organises information into the six sections displayed below, wl interconnected but can be browsed independently.

Data life cycle



RTICLE

Developing data interoperability using standards: at community use case [version 2; referees: 2 ved]

le Yeumo[1], Michael Alaux [2], Elizabeth Arnaud[3], Sophie Aubin[1], Ute Baumann[4], che[5], Laurel Cooper [6], Hanna Ćwiek-Kupczyńska[7], Robert P. Davey [8], an Fulss[9], Clement Jonquet [10,11], Marie-Angélique Laporte[3], Pierre Larmande [12,13], nier [2], Vassilis Protonotarios [14], Carmen Reverte [15], Rosemary Shrestha[9], rats[16], Aravind Venkatesan [12], Alex Whan[17], Hadi Quesneville [2]

letails

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

PLANT DATA STANDARDS : WHAT

Data standards for FAIR

Semantic

- Description of the data
- Controlled vocabularies: term name and definitions
- Ontologies: semantic links between terms
- *Biologist* driven

Structure

- Formatting and Organizing the data
- Data Models
- Standards : CSV, VCF, GFF, MIAPPE (www.miappe.org) , etc…
- *Biologist & Computer scientist* driven

Persistent Unique Identifiers

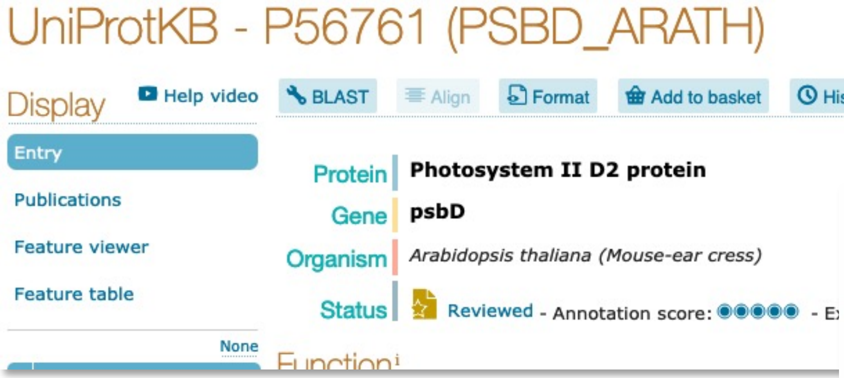
URI, gene ID, accessions ID, Trait ID, DOI,…

Technical

- Data integration and sharing
- Interoperability : tools and systems
  - GA4GH
  - Breeding API www.brapi.org
- *Computer scientist* driven

Semantic Standard: Ontologies

- Annotating one object
  - Protein, gene
  - Plant, plant anatomy, …

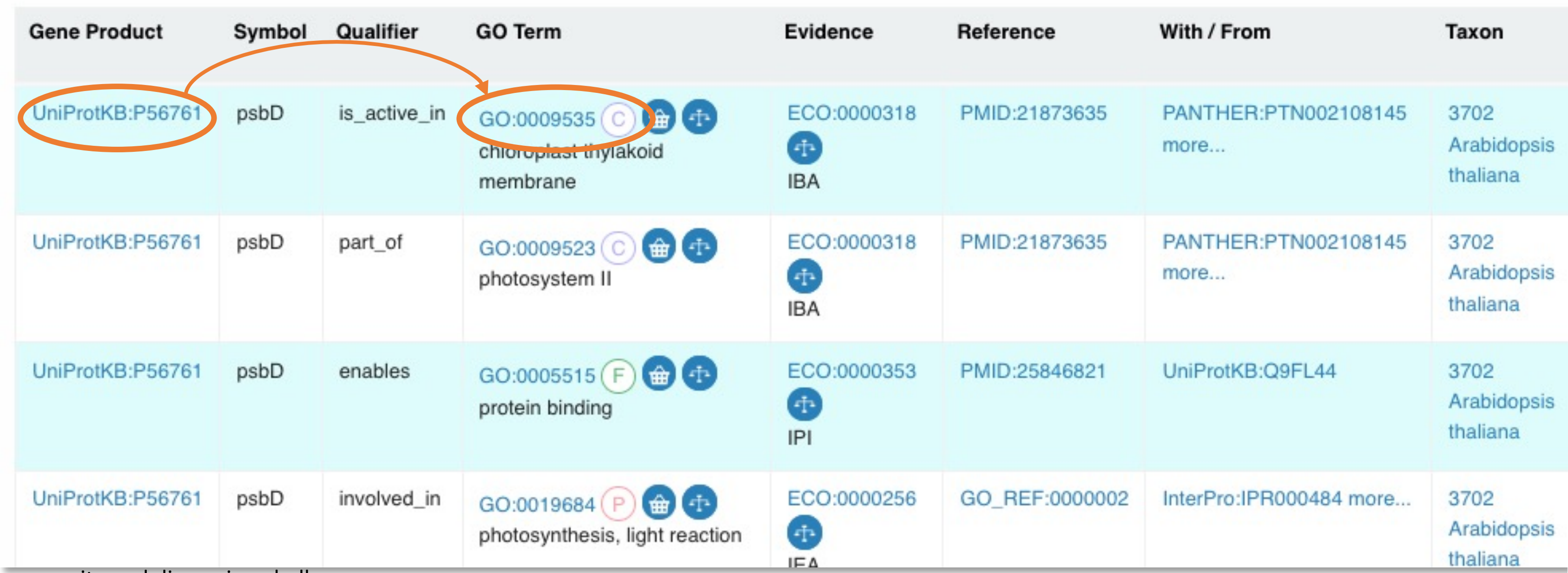


UniProtKB - P56761 (PSBD_ARATH)

Protein: Photosystem II D2 protein
Gene: psbD
Organism: *Arabidopsis thaliana (Mouse-ear cress)*
Status: Reviewed - Annotation score: ●●●●● - E

Function

- Atomic concept
  - protein function
  - cellular localization
  - …

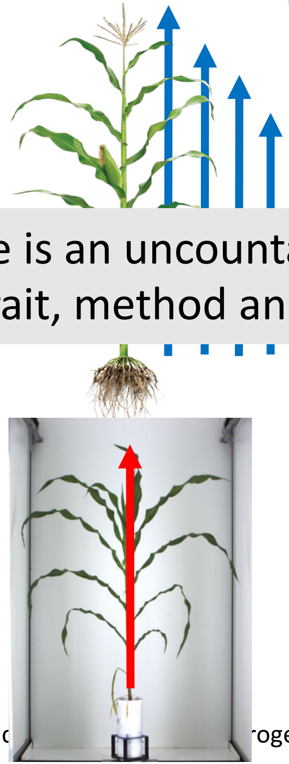


Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon
UniProtKB:P56761	psbD	is_active_in	GO:0009535 (C) chloroplast thylakoid membrane	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	part_of	GO:0009523 (C) photosystem II	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	enables	GO:0005515 (F) protein binding	ECO:0000353 IPI	PMID:25846821	UniProtKB:Q9FL44	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	involved_in	GO:0019684 (P) photosynthesis, light reaction	ECO:0000256 IEA	GO_REF:0000002	InterPro:IPR000484 more...	3702 Arabidopsis thaliana

Semantic Standard: Ontologies for Phenotype

- Describing traits/features in specific plant species
- Crop Ontology Trait + Method + Scale Semantic model

Variable identification: Plant height example



Trait + Method + Unit

T1: Plant Height

M1: Total height

M2: First tassel branch

M3: Last expanded leaf

M4: Youngest growing leaf

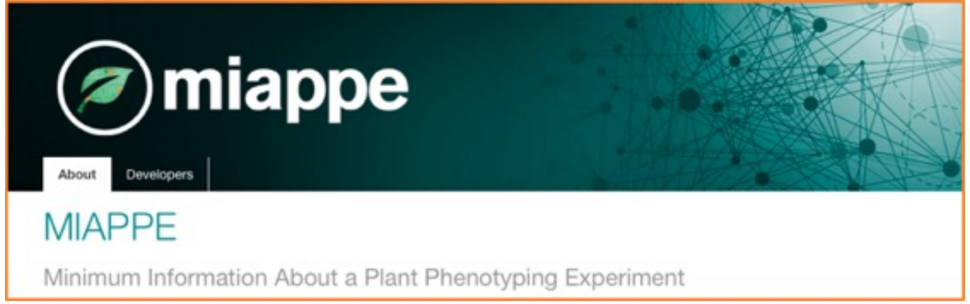
M5: Highest pixel corresponding to plant

U3: pixel

…There is an uncountable number of combinations…
Each trait, method and unit has to be identified if we want to share and reuse data

Slide from L. Cabrera-Bosquet

Phenotype Structure Standard


miappe

MIAPPE

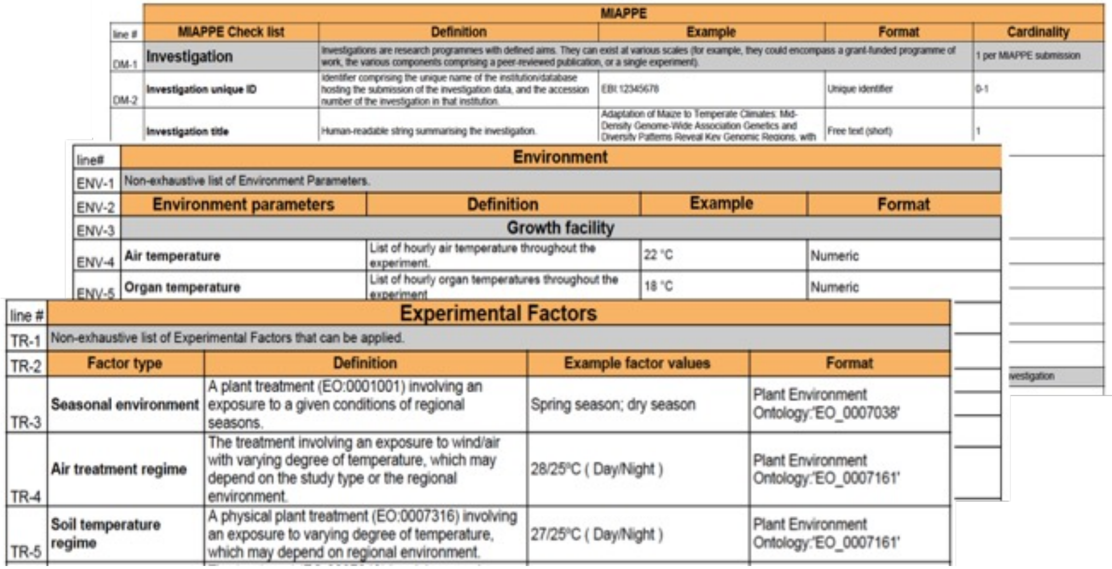
Minimum Information About a Plant Phenotyping Experiment

Minimal Information About Plant Phenotyping Experiment : version 1.1 (Jan 2019)

www.miappe.org

- Many stakeholders
  - Elixir, Emphasis, Bioversity, North American PPN
- Open Community:
  - Request for comments
  - Github Feature requests
  - Mailing lists
  - Meetings & Workgroups
- Crops and woody plants

Papoutsoglou *et al*. (2020) Enabling reusability and interoperability of plant phenomic datasets with MIAPPE 1.1. New Phytol, 227:260-273; https://doi.org/10.1111/nph.16544

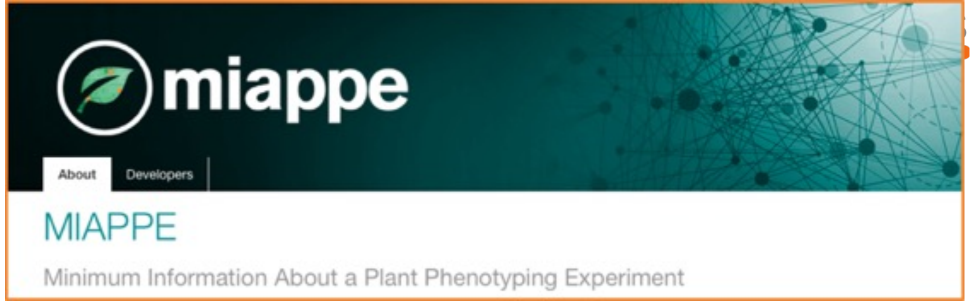


	MIAPPE				
line #	MIAPPE Check list	Definition	Example	Format	Cardinality
DM-1	Investigation	Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment).			1 per MIAPPE submission
DM-2	Investigation unique ID	Identifier comprising the unique name of the institution/database hosting the submission of the investigation data, and the accession number of the investigation in that institution.	EBI:12345678	Unique identifier	0-1
	Investigation title	Human-readable string summarising the investigation.	Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with	Free text (short)	1

line#	Environment			
ENV-1	Non-exhaustive list of Environment Parameters.			
ENV-2	Environment parameters	Definition	Example	Format
ENV-3	Growth facility			
ENV-4	Air temperature	List of hourly air temperature throughout the experiment.	22 °C	Numeric
ENV-5	Organ temperature	List of hourly organ temperatures throughout the experiment	18 °C	Numeric

line #	Experimental Factors			
TR-1	Non-exhaustive list of Experimental Factors that can be applied.			
TR-2	Factor type	Definition	Example factor values	Format
TR-3	Seasonal environment	A plant treatment (EO:0001001) involving an exposure to a given conditions of regional seasons.	Spring season; dry season	Plant Environment Ontology:'EO_0007038'
TR-4	Air treatment regime	The treatment involving an exposure to wind/air with varying degree of temperature, which may depend on the study type or the regional environment.	28/25°C (Day/Night)	Plant Environment Ontology:'EO_0007161'
TR-5	Soil temperature regime	A physical plant treatment (EO:0007316) involving an exposure to varying degree of temperature, which may depend on regional environment.	27/25°C (Day/Night)	Plant Environment Ontology:'EO_0007161'

Phenotype Structure Standard



MIAPPE

Minimum Information About a Plant Phenotyping Experiment

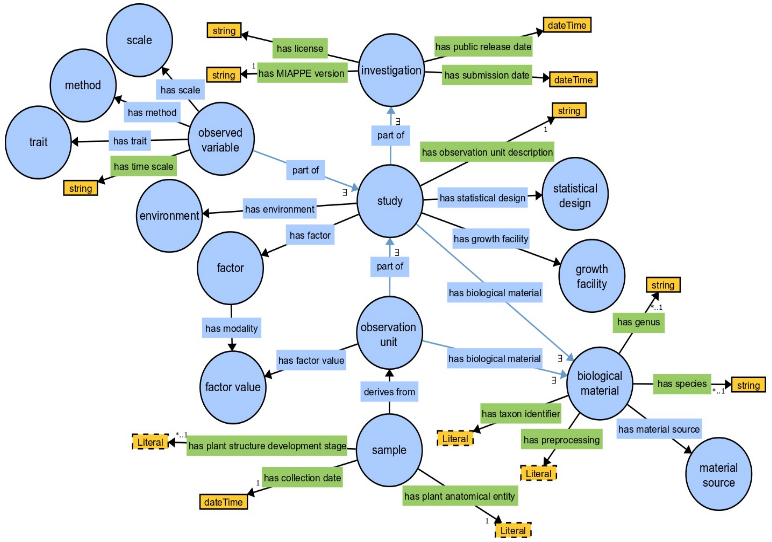
Minimum Information for Biological and Biomedical Investigations

A collection of the historical MIBBI foundry reporting guidelines. The minimum information standard is a set of guidelines for reporting data derived by relevant methods in biosciences. If followed, it ensures that the data can be easily verified, analysed and clearly

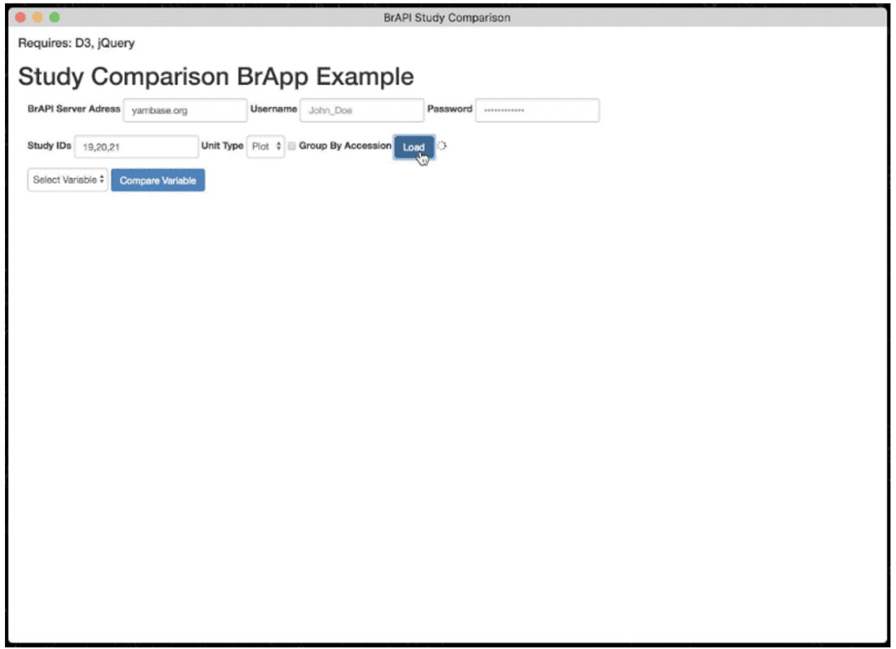
- Biologist Friendly
  - Clear definitions and examples
  - Excel templates
  - Trainings
- Minimal and sufficient list of metadata:
  - The objective of the experiment
  - Who contributed to the experiment
  - What were the experimental procedures
  - What was the biological material experimented
  - ...

Phenotype Technical Standard, MIAPPE Implementations

- Ontology, OWL Implementation
  - https://github.com/MIAPPE/MIAPPE-ontology
  - http://agroportal.lirmm.fr/ontologies/PPEO
  - Data model representation
  - Formal concepts and constraints
- File Archive
  - ISA Tab: data + metadata
  - RO Crate studies
- Web Services
  - Breeding API
  - International collaboration
  - Standard Open Web Service API
  - Information Exchange, Main target: Breeding
  - Excellence in Breeding platform (CGIAR, Peter Selby)



BrAPI



Requires: D3, jQuery

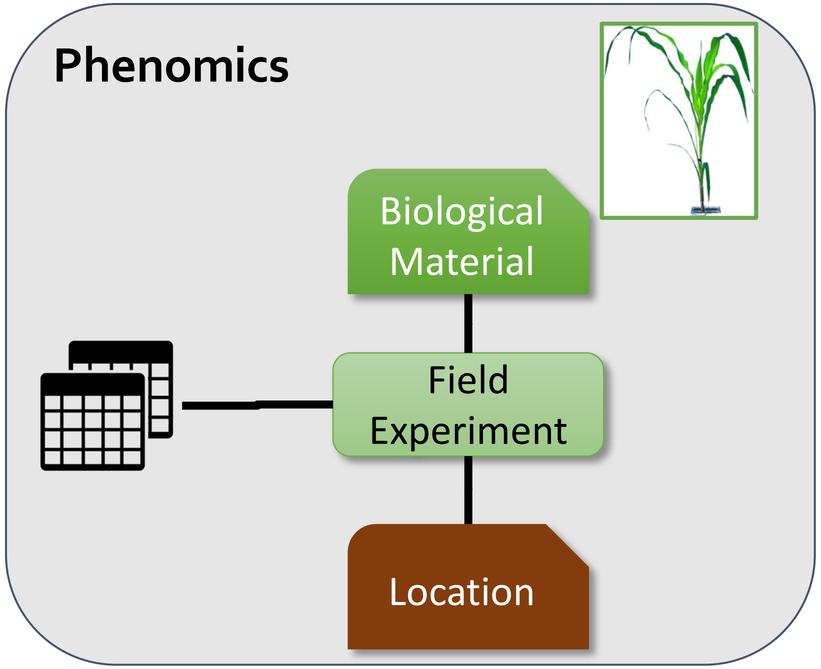
Study Comparison BrApp Example

BrAPI Server Adress yambase.org Username John_Doe Password ••••••••••••

Study IDs 19,20,21 Unit Type Plot ☐ Group By Accession

Select Variable

Data Integration between silos, From Phenotyping to Genotyping



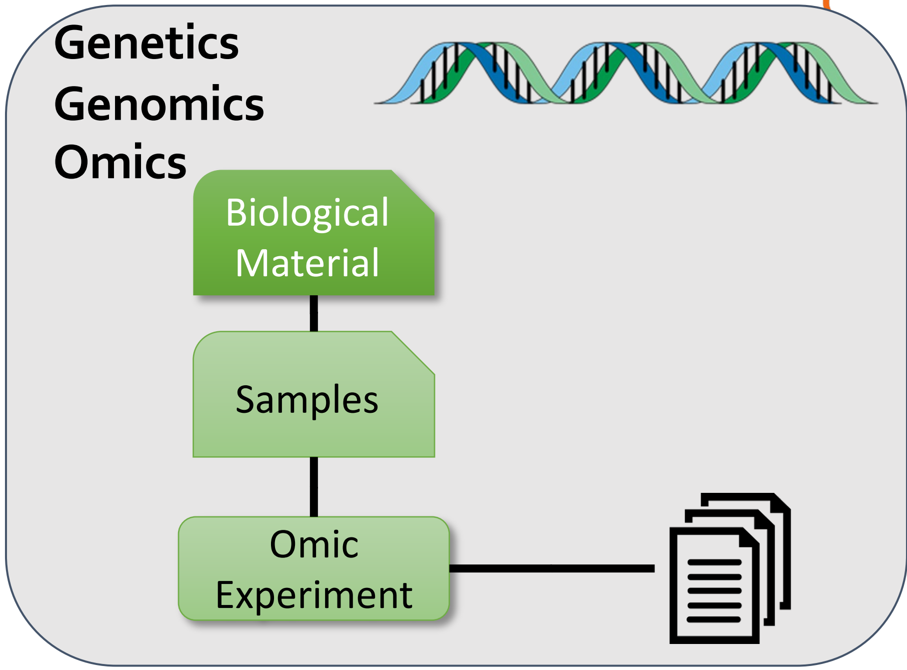
Phenomics

Biological Material

Field Experiment

Location

Identifying key resources/pivot objects



Genetics

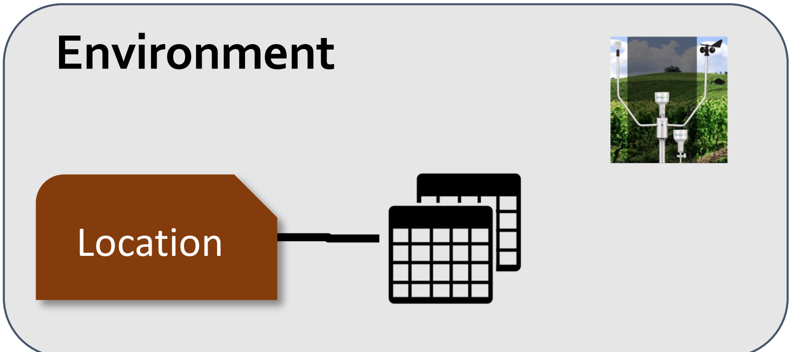
Genomics

Omics

Biological Material

Samples

Omic Experiment

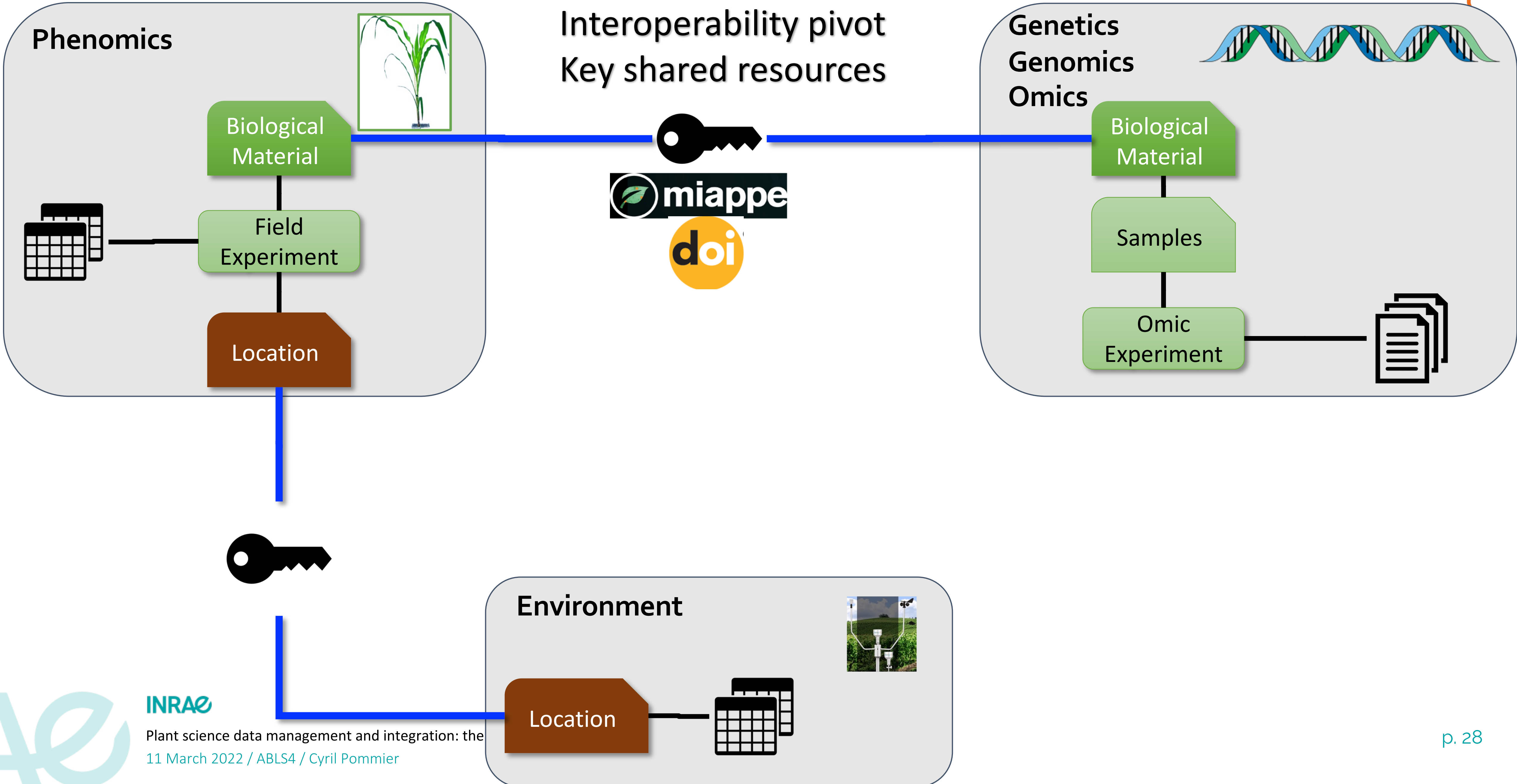


Environment

Location

Data Integration between silos, From Phenotyping to Genotyping

Interoperability pivot
Key shared resources



Phenomics

Biological Material

Field Experiment

Location

miappe

doi

Genetics
Genomics
Omics

Biological Material

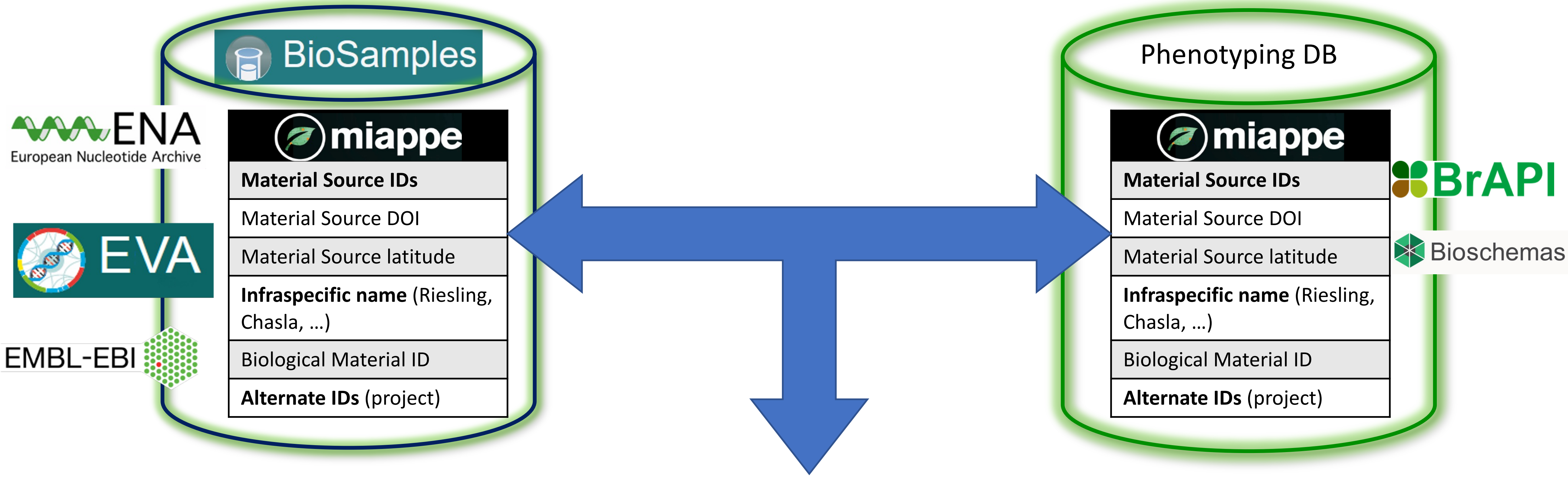
Samples

Omic Experiment

Environment

Location

Data Integration between silos, From Phenotyping to Genotyping



miappe

Material Source IDs

Material Source DOI

Material Source latitude

Infraspecific name (Riesling, Chasla, …)

Biological Material ID

Alternate IDs (project)

Phenotyping DB

miappe

Material Source IDs

Material Source DOI

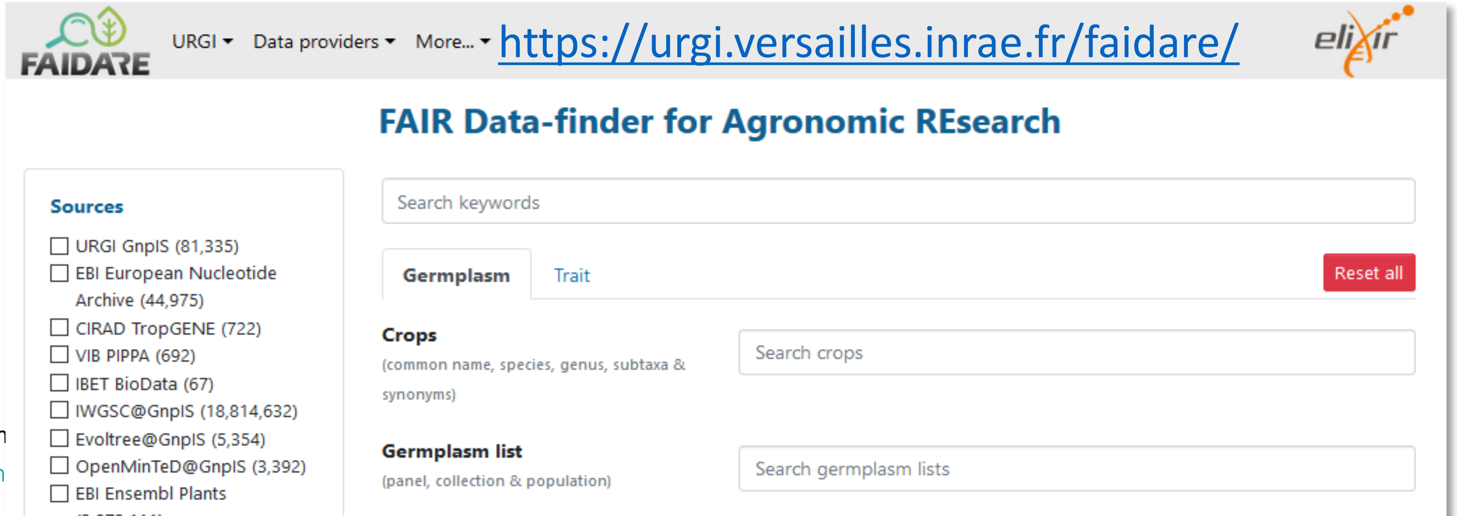
Material Source latitude

Infraspecific name (Riesling, Chasla, …)

Biological Material ID

Alternate IDs (project)

Community data discovery portals



https://urgi.versailles.inrae.fr/faidare/

FAIR Data-finder for Agronomic REsearch

Sources

- □ URGI GnpIS (81,335)
- □ EBI European Nucleotide Archive (44,975)
- □ CIRAD TropGENE (722)
- □ VIB PIPPA (692)
- □ IBET BioData (67)
- □ IWGSC@GnpIS (18,814,632)
- □ Evoltree@GnpIS (5,354)
- □ OpenMinTeD@GnpIS (3,392)
- □ EBI Ensembl Plants

Search keywords

Germplasm Trait Reset all

Crops
(common name, species, genus, subtaxa & synonyms)

Search crops

Germplasm list
(panel, collection & population)

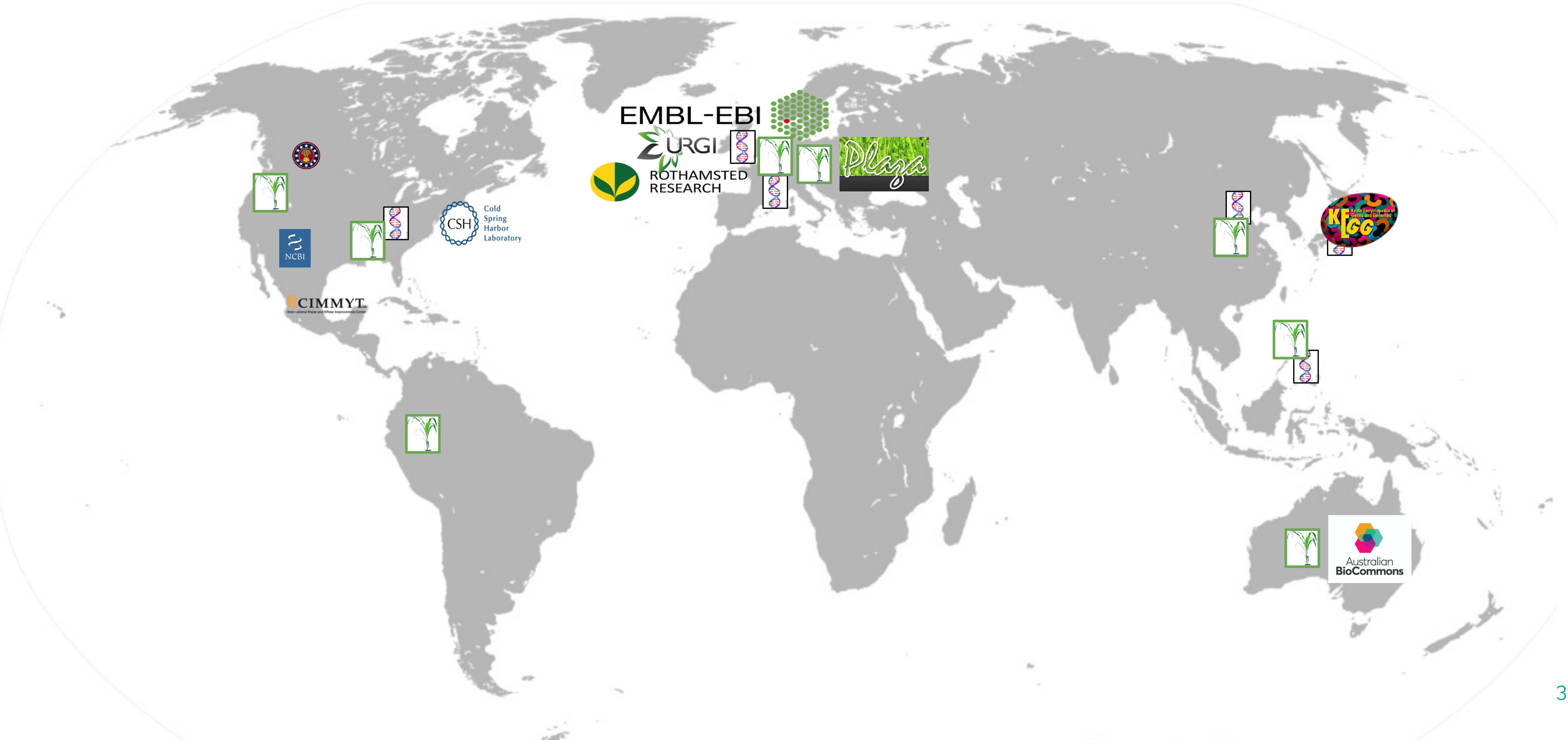
Search germplasm lists

Global Data discovery portal

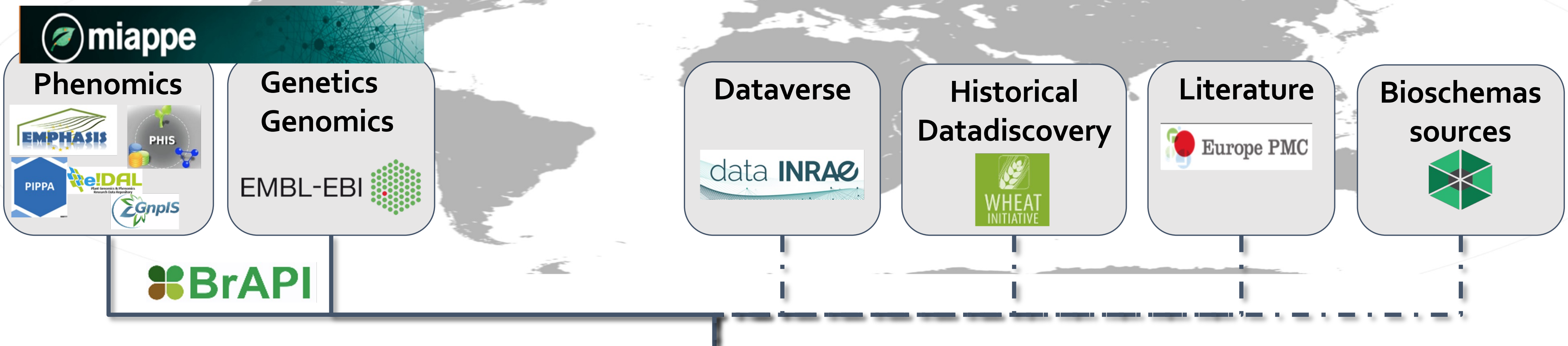
Dispersed data

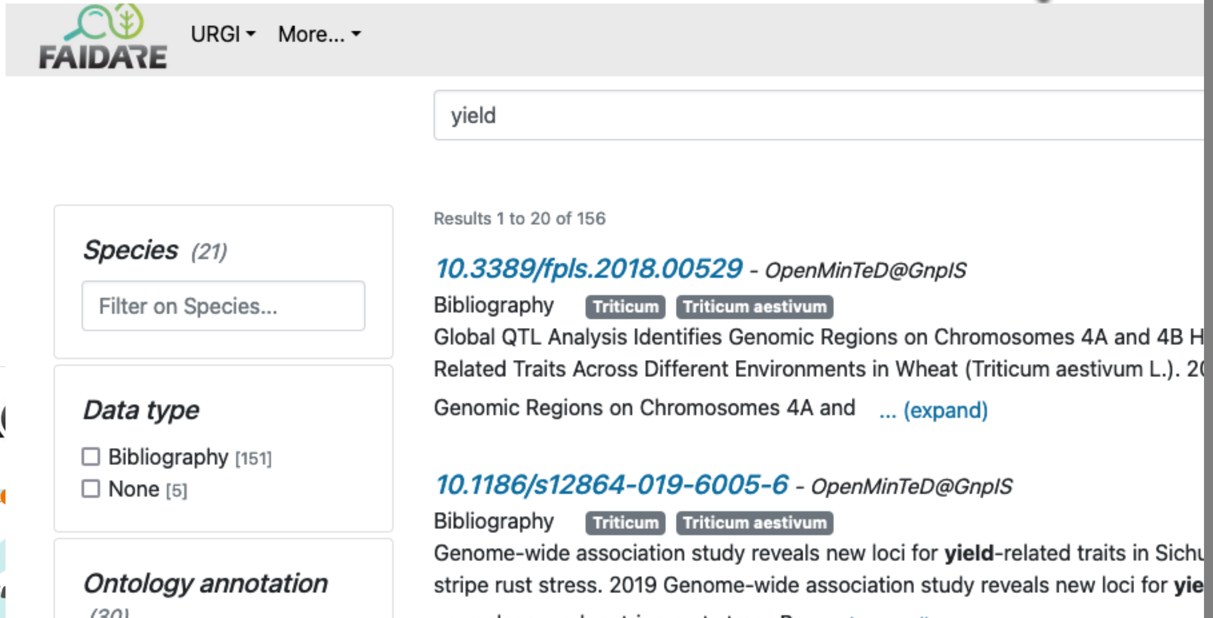
Heterogenous data

Dedicated repositories & Archives


EMBL-EBI
URGI
ROTHAMSTED RESEARCH
Plaza
CSH Cold Spring Harbor Laboratory
NCBI
CIMMYT
International Maize and Wheat Improvement Center
KEGG
Kyoto Encyclopedia of Genes and Genomes
Australian BioCommons

FAIDARE: Global Data discovery portal





yield

Species (21)

Filter on Species...

Data type

☐ Bibliography [151]
☐ None [5]

Ontology annotation
(30)

Results 1 to 20 of 156

10.3389/fpls.2018.00529 - *OpenMinTeD@GnpIS*

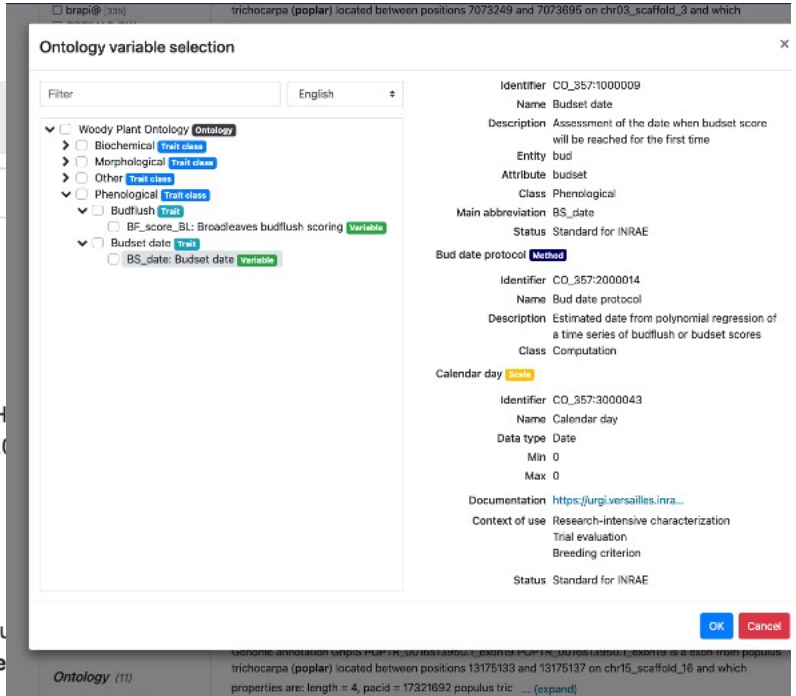
Bibliography Triticum Triticum aestivum

Global QTL Analysis Identifies Genomic Regions on Chromosomes 4A and 4B H
Related Traits Across Different Environments in Wheat (Triticum aestivum L.). 20
Genomic Regions on Chromosomes 4A and ... (expand)

10.1186/s12864-019-6005-6 - *OpenMinTeD@GnpIS*

Bibliography Triticum Triticum aestivum

Genome-wide association study reveals new loci for **yield**-related traits in Sichu
stripe rust stress. 2019 Genome-wide association study reveals new loci for **yie**



trichocarpa (**poplar**) located between positions 7073249 and 7073695 on chr03_scaffold_3 and which

Ontology variable selection

Filter

- ☐ Woody Plant Ontology Ontology
  - ☐ Biochemical Trait class
  - ☐ Morphological Trait class
  - ☐ Other Trait class
  - ☐ Phenological Trait class
    - ☐ Budflush Trait
      - ☐ BF_score_BL: Broadleaves budflush scoring Variable
    - ☐ Budset date Trait
      - ☐ BS_date: Budset date Variable

Identifier	CO_357:1000009
Name	Budset date
Description	Assessment of the date when budset score will be reached for the first time
Entity	bud
Attribute	budset
Class	Phenological
Main abbreviation	BS_date
Status	Standard for INRAE

Bud date protocol Method

Identifier	CO_357:2000014
Name	Bud date protocol
Description	Estimated date from polynomial regression of a time series of budflush or budset scores
Class	Computation

Calendar day Scale

Identifier	CO_357:3000043
Name	Calendar day
Data type	Date
Min	0
Max	0
Documentation	https://urgi.versailles.inra...
Context of use	Research-intensive characterization Trial evaluation Breeding criterion
Status	Standard for INRAE

trichocarpa (poplar) located between positions 13175133 and 13175137 on chr15_scaffold_16 and which
properties are: length = 4, pacid = 17321692 populus tric ... (expand)

Full text

+

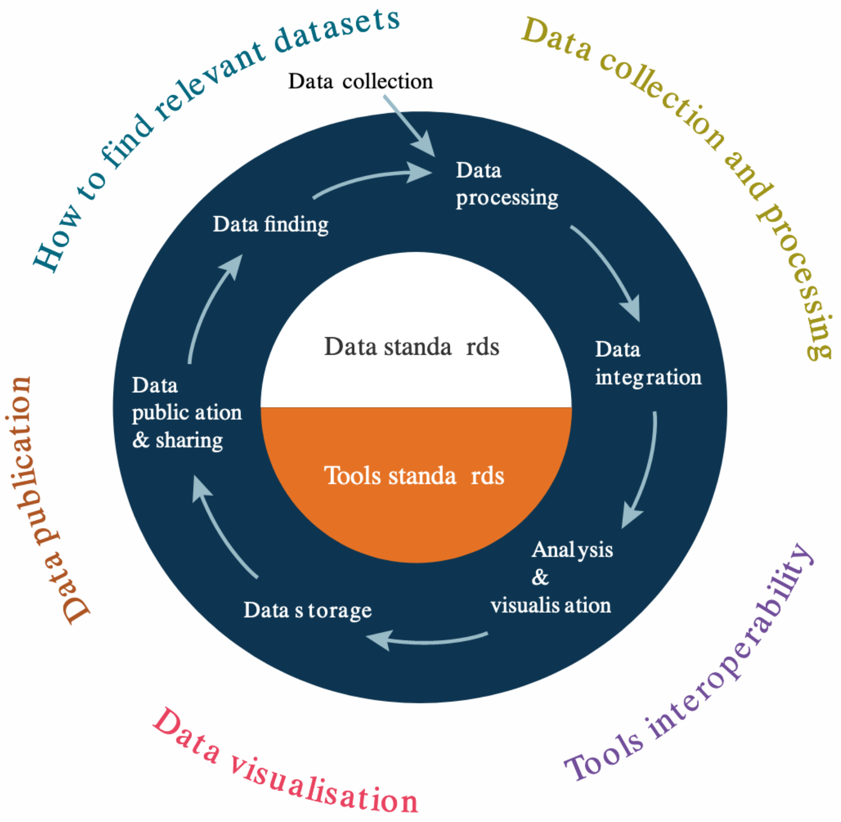
Fine criteria

+

Link back

Take Home Message

- Data integration relies on a complex lifecycle
- Both:
  - Deriving and reducing data
  - Linking different datasets
- All steps must be defined
  - → data management plan
- Not all step of data must be shared
- Raw and final data should be shared
- With sufficient provenance
- Open science (policy): Publication and Findability are keys (Data tombs effect)



How to find relevant datasets

Data collection and processing

Tools interoperability

Data visualisation

Data publication

Data collection

Data processing

Data integration

Analysis & visualisation

Data storage

Data publication & sharing

Data finding

Data standards

Tools standards

Aknowledgments

Elixir Plant community & platforms

Beier S., Gruden C., Pommier C., Coppens F, Scholz U, Lange M., Contreras B., Adam Blondon AF, Faria D, Chavez I, Miguel C, Droedsbek B, Finkers R, Papoutsoglou E, Olster R, Ramsak Z, …

H2020 AGENT

N. Stein (IPK, coord), P. Kersey (RBGK), M. Alaux (INRAE), S. Weise (IPK), C. Pommier (INRAE), M. Lange (IPK), R. Finkers (WUR), *J. Destin (INRAE)*

Crop Ontology

Arnaud E, Laporte MA, …

MIAPPE community

ELIXIR Plant Community,

Krajewsky P, Cwiek H, Tardieu F, Usadel B, Arend D, Arnaud E, Junker A, King G, Laporte MA, Poorter H, Reif J, Rocca-Serra P, Sansone SA, Kersey P,

And many more!

Breeding API

Selby P, Mueller L, Robbins K, Backlund JE, … ,

And many more!

Emphasis

Tardieu F, Usadel B, Arend D, Junker A, Poorter H, Neveu P, Pierushka R, Shur U…

And many more!